知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

基于关键词挖掘的热线文本数据犯罪线索筛查方法研究

甄沐华[1] 陈鹏[1] 王坤[2] 范子杨[1] 王者[1]
1. 中国人民公安大学信息网络安全学院 北京 100038
2. 济南市公安局 济南 250099

摘要：[目的 / 意义] 针对公安业务中对热线文本数据犯罪线索关键信息识别与筛查时存在的信息化分析能力不足问题，提出一种基于关键词挖掘的热线文本数据犯罪线索筛查方法，帮助业务部门提高相关情报研判效率，使得犯罪线索筛查工作更加信息化和科学化。[方法 / 过程] 考虑到直接采用文本类等算法方法或因有效信息样本量占比过小使得模型训练不充分，本文首先对已知犯罪线索进行基于文本相似度的种子词集抽取，然后采用 Word2Vec 对种子词汇从同类词、替代词两个角度扩展构成专业词库，最后使用基于语义的积分筛查模型实现对热线文本数据中犯罪线索筛查。[结果 / 结论] 对济南市 1 050 条先验热线文本数据作犯罪线索筛查实验，并进行实际比对与结果指标分析，得到结果召回率 86%，可以认为本文所述基于语义的积分筛查方法对济南市热线文本数据内犯罪信息具体性识别达到预期效果并实现犯罪线索有效筛查。

关键词：热线文本 专业词库 文本相似度 犯罪线索筛查
分类号：TP391；G250

引用格式：甄沐华 , 陈鹏 , 王坤 , 等 . 基于关键词挖掘的热线文本数据犯罪线索筛查方法研究 [J/OL]. 知识管理论坛 , 2022, 7(5): 539-548[引用日期]. http://www.kmf.ac.cn/p/313/.

1 引言

电话热线是便民服务的重要举措，同时，热线文本数据往往因潜藏着一些犯罪线索（指可供侦查、调查和控制的有关犯罪活动的情报信息）而成为公安机关犯罪线索排查的重要数据来源。目前，公安机关在处理热线文本数据时，多采用“标签体系 + 人工筛查”的方法，即执法人员首先通过分类标签定位至可能出现

基金项目：本文系北京市自然科学基金项目“数据驱动下的城市犯罪风险机理分析与防控优化研究”（项目编号：9192022）研究成果之一。
作者简介：甄沐华，硕士研究生；陈鹏，副教授，博士，通信作者，E-mail：chenpeng@ppsuc.edu.cn；王坤，食药环侦支队一大队大队长；范子杨，本科生；王者，本科生。
收稿日期：2022-07-01 发表日期：2022-09-30 本文责任编辑：刘远颖

犯罪关键信息的数据类目，再快速浏览数据详情内容字段并根据经验知识识别事件关键信息，最后研判该数据是否作为犯罪线索输出。但由于详情内容字段数据多以大段落文本形式呈现，且其中包含的有效关键信息词汇单元占比较小，在提取和挖掘关键信息时具有相当的困难，使得传统人工筛查模式中存在有效研判效率较低、数据利用不充足等问题[1-2]。

热线文本数据犯罪线索筛查工作的关键在于对数据文本内容中代表犯罪语义关键信息的识别和提取。目前，在文本内容关键信息抽取方面，研究人员进行了大量的研究，其中基于词频的关键词提取（TF-IDF、LDA 等）是一种常用的方法，但是当关键信息词汇单元数量在文本中占比较小时，基于词频的关键词提取方法不能够满足文本分析的需求，与此同时，在中文文本分析时，基于词频提取的关键词还存在着语义歧义问题[3]。对此，一些研究人员提出通过词向量技术（Word2Vec）构建关键信息词库，结合关键词抽取、文本相似度计算等文本分析方法以解决关键信息词汇单元占比小及语义歧义问题对文本分析的影响。例如，彭云等利用基于语义关系约束的 SRC-LDA 主题模型对商品评论文本进行了主题词提取，实现了对商品评论主题词的有效提取[4]；刘耕等利用关联词和 Jaccard 系数扩展规则设计了敏感词库并对网络舆情敏感文本进行了敏感信息检索和提取，实现了网络敏感信息可靠率 10% 以上的提升[5]；刘亚桥等利用词向量模型构建的摄影领域评论情感词典对摄影评论数据进行了摄影情感信息提取并做进一步语料分类，实现了基于情感词典下对摄影领域评论语料分类[6]；谭敏博等对谷类作物病害数据进行了谷类作物病害特征信息提取，实现了对谷类作物病害特征属性识别的精准查询[7]；夏松等利用基于 Word2Vec 技术的语义近似匹配对微博类社交平台短文本构建了网络谣言敏感词库，实现了基于网络谣言敏感词库的网络谣言有效识别[8]；唐晓波等联合 TF-IDF 方法与词向量特征扩展方法对医疗问答社区健康问句短文本提取了健康信息关键词并集合作为健康问句关键信息词库，实现了基于健康问句关键信息词库的健康问句文本的有效分类[9]；姜天宇等利用词向量构建和 TF-IDF 加权方法对新华社不同类别邮件进行了邮件主题信息关键词提取，进一步结合改进的朴素贝叶斯树方法实现了对新华社邮件的文本分类[10]。

从目前研究进展来看，关键词、特征词提取等自然语言处理技术已在新闻学等诸多领域得到了应用，并达到了较好的应用效果。但在当前，各类公安业务处理线索数据文本工作时受限于信息表达规范化不足、有效信息分散等问题而仍采用传统人工筛查模式，缺少针对特定类型犯罪线索的有效信息化挖掘方法。为此，本文以热线文本数据为例，立足犯罪线索文本特点，设计了抽取其中犯罪线索关键信息的方法，并根据公安机关情报研判逻辑设计了基于语义的积分筛查模型[11]，从而提升公安机关文本数据中信息化获取犯罪线索的能力。

2 关键词抽取

在“标签体系 + 人工提取”筛查方法广泛、成熟的应用背景下，热线文本数据同样根据事件所涉政府业务领域不同而被赋予以业务领域相应粗粒度标签，事件详情内容则不做标签处理。而热线文本数据中的犯罪线索往往从事件详情内容字段中挖掘分析得到，且代表犯罪线索语义的关键信息在详情内容文本中位置分散、数量较其他信息占比小、不具有明显文本句式结构化特征，常见表达形式有单词汇表达、短语句式表达两种，如“侦查”“予以 / 取缔”。与此同时，构建专业词库时不可避免地对短语句式进行再分词处理，若采用文本类等自动化算法直接对文本进行处理，则再分词后存在的大量无独立语义词汇将对结果准确性有明显影响。

目前，公安机关民警对热线数据中犯罪线索的排查和识别主要通过关键词来进行判定，但由于来电人表达方式和习惯的不同，一些涉

ChinaXiv合作期刊
知识管理论坛，2022（5）:539-548
DOI: 10.13266/j.issn.2095-5472.2022.044

嫌犯罪的表述可能存在着句式结构和语义歧义等问题。因此，要尽可能地达到对热线文本数据中犯罪线索的排查和识别，首先需要确定数据中已有的代表犯罪语义关键信息词汇（种子词集），并在此基础上关联相关的同义词和近义词（扩展词集），最终实现热线文本数据犯罪线索的关键词库的构建。

2.1 种子词集构建

词向量技术（Word2Vec）是一种基于上下文分布表示词义的技术方法，其专注于无标注数据，利用神经网络语言模型从大量文本中学习语义信息。词向量技术常常用于计算词语间、句子间或者其他长文本间的相似度，并具有良好效果[12-16]。

在种子词集构建上，本文首先收集执法部门的犯罪信息词汇作为经验知识词集，随后以全量数据语料作为训练语料库，得到全量数据Word2Vec词向量模型、已知属性（普通事件/疑似犯罪线索事件）的数据语料Word2Vec词向量、经验知识词集基于全量语料上下文语义的词向量，最后，以已知属性数据语料词向量作为种子词集识别抽取的数据基础，以经验知识词汇集词向量为对照匹配变量集，通过向量间映射计算得到二者文本相似度，实现对已知属性数据中符合相似度要求的信息词汇抽取并集合得到种子词集，其流程如图1所示：

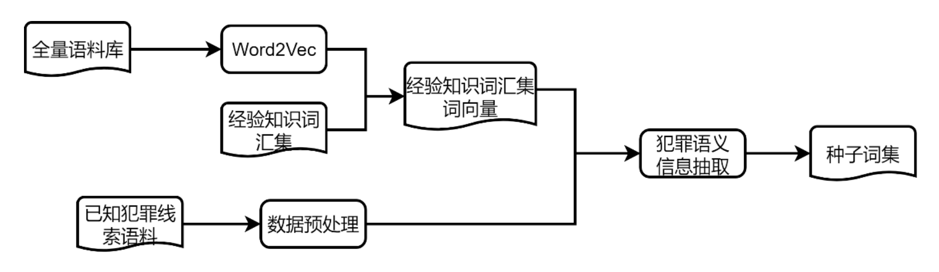


图1 构建种子词集流程

抽取得到的种子词汇分为两类：代表疑似犯罪线索事件语义的词汇 *Word_T*（下同），代表普通事件语义的词汇 *Word_F*（下同）。此处所指"疑似犯罪线索事件"即可根据相关法律规定属于公安机关犯罪活动侦办的事件，包括可判定为有违法行为但未达犯罪标准的、需要进一步确认的及已立案需督办的线索事件；普通事件即根据相关法律规定不属于公安机关管辖的事件，包括经有关办理单位确认后反馈为恶意、重复拨打的无效热线事件。

为确定抽取得到的种子词汇在犯罪线索筛查中的可靠性，通过回溯已知属性数据本身，定义回溯值为某种子词汇所属数据属性为犯罪线索的数据数量（回溯数）与其在全量数据中出现次数（词频）的比值，代表了该词汇在犯罪线索筛查过程中的可靠性，公式（1）：

$$P_{(word)} = \frac{n_{(word)}}{N_{(word)}} \qquad 公式（1）$$

其中，$P_{(word)}$ 代表种子词汇回溯值，$n_{(word)}$ 代表种子词汇回溯数，$N_{(word)}$ 代表种子词汇在全量数据中词频。将得到的回溯值作为对应种子词汇在犯罪线索筛查模型中的权重系数。

2.2 扩展词集

考虑到同一语义的表达会以不同的词汇和句式结构呈现，为了实现专业词库的有效覆盖和扩展，从种子词集的同类词、替代词2个方面进行词库的扩展，另结合舆情领域公开敏感词库共同构成扩展词集。扩展词集的词汇可靠性由扩展词汇与种子词汇的文本字面距离相似度确定，本文采用余弦距离相似度（Cosine Similarity）计算得到，如公式（2）所示：

$$Similarity(\omega_1, \omega_2) = \cos\theta = \frac{\omega_1\omega_2}{|\omega_1||\omega_2|} \qquad 公式（2）$$

对于同类词集扩展，Word2Vec方法计算所得词向量能够反映出词汇所处上下文和语义关系。首先通过全量语料的Word2Vec词向量模型

得到种子词集的词向量，再以全量数据语料库构建的 Word2Vec 词向量模型为同类词集识别抽取的数据基础，以种子词集词向量对照匹配变量集，计算得到二者文本相似度，实现在全量语料库中基于上下文语义关系的关键信息同类词汇的抽取，并将相似度作为对应词汇在犯罪线索筛查模型中的权重系数，其流程如图 2 所示：

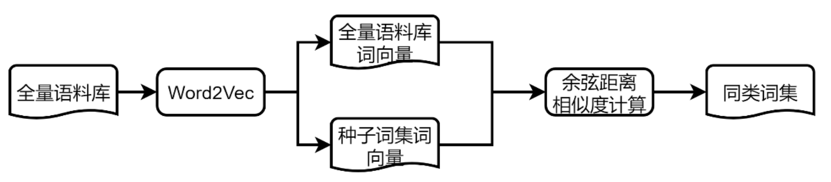


图 2 构建同类词集流程

对于替代词集扩展，考虑到同一语义可由不同词汇表达，以种子词集在中文表达中的近义词作为其替代词。利用种子词集基于全量语料的 Word2Vec 词向量模型的词向量，结合近义词查找工具，在以开源维基百科中文语料库中寻找近义词并计算二者文本相似度，实现基于公开中文语料库的关键信息替代词汇的抽取，将相似度作为对应词汇在犯罪线索筛查模型中的权重系数，其流程图如图 3 所示：



图 3 构建替代词集流程

3 犯罪线索筛查模型

3.1 犯罪线索积分预警模型

积分预警模型是基于大数据背景下的情报主导警务模式应运而生的公安数据挖掘手段[11]。该模型方法以某事件发生为预警对象，将可能影响该事件发生的因素罗列出来，并按照因素的影响性程度赋予相应的权重分值，每当某个因素出现时，都会以和的形式计算出相应分值，直到所有的因素都被积分出来。积分分值代表事件发生的定量描述，可表示为：

$$Y=\sum_{i=1}^{n} y_i p_i \qquad \text{公式（3）}$$

其中，i 为影响因素，y 为分值设定，p 为该因素权重系数。针对本文研究热线数据，单条待筛查数据积分总值由其与各类型词集匹配后产生的各积分值构成。各类型词集积分值由分属两个不同属性的词集积分值构成。各个词集的积分值影响因素为符合条件的单一词汇的相似度、该单一词汇权重值及与词集词汇匹配到相同词汇的个数。除此之外，舆情领域公开敏感词集只做相同词汇计数积分处理。单条数据于词集的积分值计算规则如下：

$$S_{(dic)} = aS_{(Word_T)} + bS_{(Word_F)} \qquad \text{公式（4）}$$

$$SUM_{(data)} = \sum_{dic} S_{(dic)} + Counts_{(internet)} \qquad \text{公式（5）}$$

其中，$S_{(dic)}$ 代表某类型词集（种子词集、同类词集、替代词集）的积分值，$S_{(Word_T)}$ 及 $S_{(Word_F)}$ 代表某类词集中代表疑似犯罪线索事件语义的词集（T）或代表普通事件语义的词集（F），a、b 为该词集的权重系数，SUM 代表总积分结果，$Counts_{(internet)}$ 代表匹配过程中出现的舆情领域公开敏感词集中不重复计数的词汇数目。

3.2 犯罪线索筛查算法

在采用“标签体系”对数据已做粗粒度分类背景下，本文研究文本数据中包含事件详情内容信息和标点符号、语气词等无效信息。据此，在匹配筛查之前需要对待筛查数据作预处理：通过中文分词工具 Jieba 对待筛查数据进行分词处理，为避免分词粒度不同造成后续匹配失败，在精确分词模式基础上设计自定义分词标准；对分词后数据，使用自定义停用词表去除标点符号、干扰词等无效文本。

本文采用基于语义的积分预警模型实现对热线文本中犯罪线索筛查，即专业词库中词汇的可靠度（权重值）与匹配时的文本相似度共同控制筛查结果。对于单条待筛查数据，筛查流程主要从 3 个层次循序进行：待筛查数据词

DOI: 10.13266/j.issn.2095-5472.2022.044

汇与某词集词汇匹配相似值计算、单条数据与专业词库中某词集匹配积分值运算、单条数据与专业词库积分值运算。

对于待筛查数据词汇与某词集词汇匹配相似值计算（match(*seg*,*word*)），即单条待筛查数据中某词汇（*seg*）与专业词库中某词集中某词汇（*word*）的相似值计算，具体步骤如下：①判断输入的两词汇是否相同，若相同则相似值记为 1，否则进行②；②判断两词汇是否同时存在于已训练好的 Word2Vec 词向量模型中，若存在则计算两词汇文本相似度后进行④，否则进行③；③在基于维基百科语料的词向量模型中得到 *seg* 的词向量，进而计算两词汇文本相似度，后进行④；④判断相似度是否大于或等于设定阈值，若满足则记录该相似度，否则结束本次相似值计算；⑤将记录的二者文本相似度与本次匹配的 *word* 对应权重值 *p* 作乘积运算，结果作为两词汇的相似值。

对于单条待筛查数据与某词集的相似值运算（sim(*data*,*dic*)），以分词后的待筛查数据、专业词库中某词集作为输入项目。遍历输入数据集合中元素并做碰撞匹配，结合 match(*seg*,*word*) 模块，对每次遍历产生相似值作求和运算。与此同时，计算某词集中词汇在待筛查数据中相同个数，再将求和运算结果与词汇相同个数求和得到该待筛查数据与某词集的相似值。

对于单条待筛查数据与专业词库相似值积分运算与结果输出（sim(*data*,*all*)），待筛查数据经上述处理后，分别得到该待筛查数据与所有词集的相似值。根据 2.1 设计的积分运算规则计算该条数据与专业词库相似值积分运算结果并输出。单轮待筛查数据集筛查完成后，可将此轮数据加入数据库中实现数据动态更新。

4 实验验证

4.1 数据来源及示例

本文主要采用依托于 Python3.0 编程语言环境的 gensim.Word2Vec 词向量模型工具构建 Word2Vec 词向量模型。实验数据来源于济南市公安局食药环支队提供的 12345 市长热线数据；时间跨度为 2020 年 1 月至 2021 年 3 月；数据分别涉及食药安全、医药监督、环境保护、疫苗注射 4 个领域，共 8 万多条；参考实际公安工作业务流程，研究数据字段为已由相关行政单位核实的热线事件回复内容，旨在发现线索、督办线索，实验数据语料部分示例及数据属性如表 1 所示：

表 1 实验语料部分示例

序号	事件文本内容（脱敏处理）	数据属性
1	有人在**社区东北角*私自*打了两口机井，据有村民反映村书记赵*与别人合伙，打算从其他地区运输*有害气体*投放谋取利益。来电投诉，要求*追究责任*。	犯罪线索（线索发现）
2	收到*先生诉求后，派出所所领导*高度重视*，原*局的*督办*件中了解到，举报**厂排放重金属，此项*超出*派出所*管辖范围*，系环保部门管辖，关于**厂经理马**逃避故意杀人、投毒、非法经营、污染环境问题，经过调阅档案，*未查询到*此案件，电话联系*先生，*先生强烈要求检察院处理，不要求公安机关处理，申请*不计入考核*。	普通事件
3	市生态环境保护综合行政执法支队执法人员就举报人反映的问题对**公司莱芜分公司进行了现场*核查*，具体核查情况是……。现场*检查*时，**公司各单位*正常*生产，污染防治设施均*正常*运行，*未发现*有在线数据超标的现象。执法人员现场要求**公司各单位严格生产管理，确保各项污染物达标排放，同时要求加强对区管企业的监管力度。关于**厂污染的投诉问题，我局已按照*信访*程序办理。建议12345热线*不列入年终满意率考核*，谢谢！	普通事件
4	针对“网传济南某整形机构女老板殴打顾客的视频”一事，经济南市公安局高新分局*立案侦查*，*犯罪嫌疑人*刘某明*涉嫌*非法拘禁罪，被依法*刑事拘留*；*嫌疑人*曲某、孙某笑*涉嫌*非法拘禁罪，被依法*取保候审*。案件正在进一步*侦办*中，*调查处理*情况将及时向社会公布。	犯罪线索（线索督办）
……	……	……

4.2 专业词库构建

4.2.1 种子词集

根据 1.1 所述种子词集构建方法，通过遍历学习集中经验知识词汇，对预处理后的已知属性数据采用 Word2Vec 词向量工具与经验知识词集中词汇作文本相似度计算，将相似度高的词汇集合，并入经验知识词集后作为种子词集。基于不同属性的数据得到种子词集分为两类：以 seed_T 指代代表疑似犯罪信息语义的词集，以 seed_F 指代普通事件信息语义的词集。实验中，共得到 94 个种子词汇，如表 2，其中 seed_T 词集 55 个，seed_F 词集 39 个。

表 2 种子词集词汇部分示例

类型	词集示例
seed_T	已/立案、口头劝诫、勒令、实施/扣押、涉嫌违法、正在/调查处理、采取措施、限期整改、线索已移交、取缔、依法处理、情节严重、落实查处……
seed_F	不/纳入/考核、不予/立案、恳请、恶意/举报、继续/监督、没有/发现、不/存在、不/属实、正常/现象、达成一致、自行协商/解决、不再/追究、不予/受理……

进一步地，对生成的种子词汇通过公式（1）并结合分层抽样方法计算词汇回溯值。图 4 为 seed_T 词频和回溯数关系图，图 5 为 seed_T 回溯值趋势图。对于 seed_T 中词汇，词汇的回溯数在词频占比中呈现明显不均衡态势，回溯值与词频关系以无规律波动呈现。整体来看，回溯值与词频无明显伴随关系，但是各词汇回溯数与词频占比体现了犯罪信息在文本中占比小的特点。分析可知，由于 seed_T 中疑似犯罪语义多为短语句式，分词后存在 3 种性质词汇，根据词频的排序为：连词（如“已经”）、中立语义词汇（如“拍照”“调查”）、术语词汇（如“取证”“嫌疑人”）。此 3 种词汇共同作用于对文本中犯罪信息的判定，连词和中立语义词汇单独出现时难以判断语义性质且常与不同的术语词汇搭配出现，而术语词汇单独出现时则需要结合语境判断是否为犯罪语义，以词频作为犯罪线索关键信息识别标准会对结果有较大影响。图 6 为 seed_F 词频和回溯数关系图，图 7 为 seed_F 回溯值趋势图。对于 seed_F 词集，回溯数与词频成正比，也即 $n_{(word)}N_{(word)}$，回溯值趋于稳定，多集中于区间 [0.8,1)。与 seed_T 中短语形式信息不同，seed_F 为表达普通事件语义的信息，其短语形式的信息（如“不 / 列入 / 考核”“超出 / 管辖范围”）大多由否定性连词词汇和术语构成，当二者同时出现将该条数据判定为普通事件的概率几乎为 1，即具有独判性。同时，多数具有否定性的术语词汇亦具有独判性（如“驳回”“恶意投诉”），因此，一些否定性质词汇的回溯值会接近于 1，也即依据该词汇判定数据为非犯罪线索可靠性极高。图 8 展示了种子词集中分词后各字词有向网络

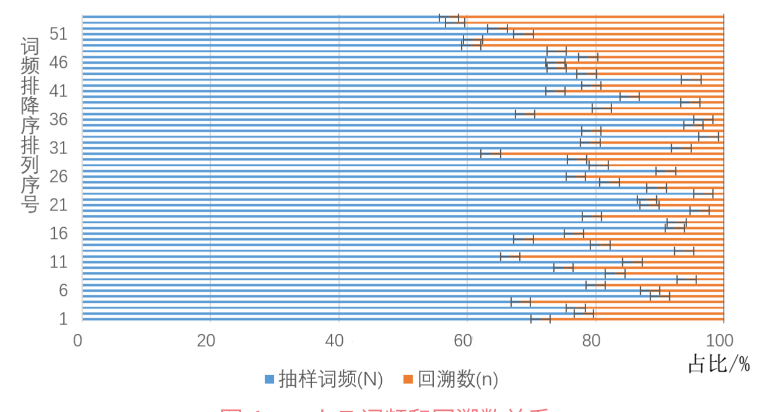


图 4 seed_T 词频和回溯数关系

知识管理论坛，2022（5）:539-548
DOI: 10.13266/j.issn.2095-5472.2022.044

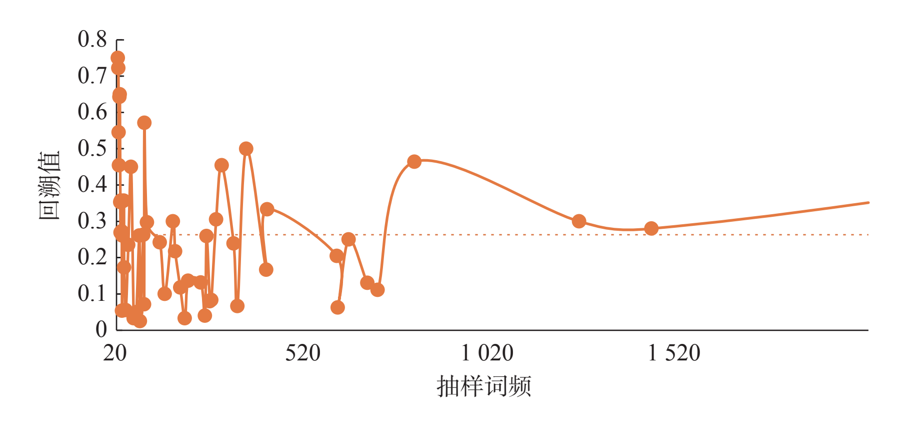


图 5 seed_T 回溯值趋势

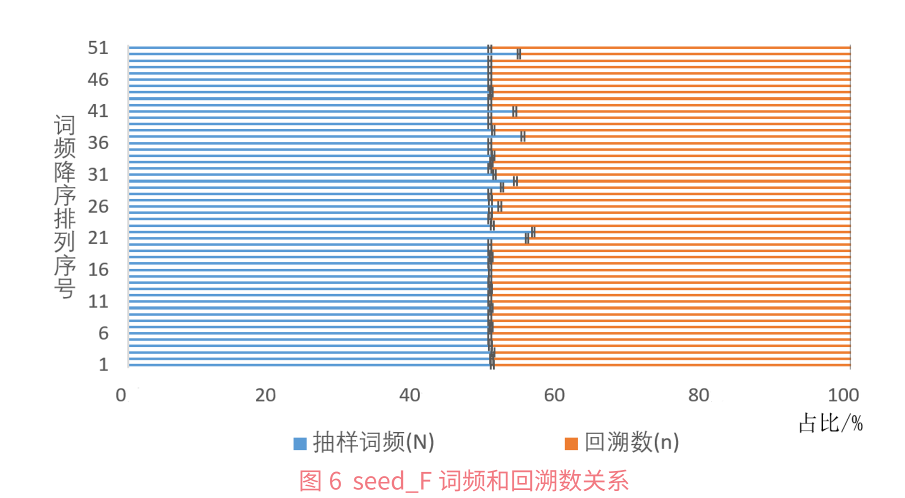


图 6 seed_F 词频和回溯数关系

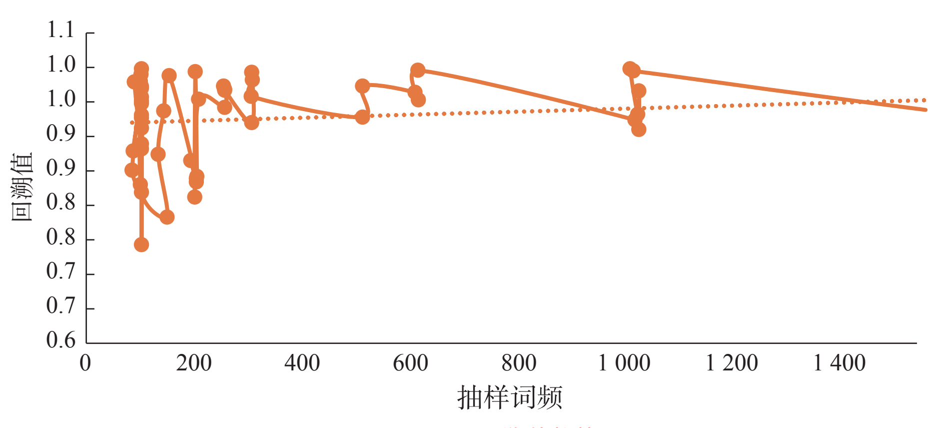


图 7 seed_F 回溯值趋势

关系图，以各字词作为节点，节点大小由词频确定，带有箭头的节点间连边为词汇组成短语的句式结构联系，边长由词汇的回溯数确定。可以发现，图中较大节点为词性是连词或语义中立性质的词汇，进一步说明了此两类词汇的可靠性较低；反之，能够明确表达疑似犯罪语义的词汇在图中表现为较小的节点，句式结构多与较大节点词汇联系，说明此类词汇的可靠

性较高。本文以字词的回溯值为其在积分筛查模型中的影响因子权重系数，能够缩小使用字词一致规则或词频系数规则作为筛查识别标准时出现结果误差。

seed_T

seed_F

图 8 seed_T 与 seed_F 中字词有向网络关系

ChinaXiv合作期刊
知识管理论坛，2022（5）:539-548
DOI: 10.13266/j.issn.2095-5472.2022.044

4.2.2 扩展词集

同类词集扩展。利用 Word2Vec 工具得到 94 个种子词集在全量语料库中的均值向量，继而通过文本相似度计算得到种子词集在全量语料库中的同类词集。实验中共取到与种子词集同类的 480 个词作为扩展的同类词，如表 3，其中 seed_T 同类词集（seed_T_similar，下同）中 251 个，seed_F 同类词集（seed_F_similar，下同）中 229 个：

表 3 同类词集词汇部分示例

类型	词集示例
seed_T_similar	案件、法定程序、立案查处、行政处罚、介入、调查、核查、核实、查处、实地调查、口头、劝诫、警告、教育批评、逾期……
seed_F_ similar	不计、考核、记入、不记、满意率、考核成绩、纳入、不列入、列为、不予、拒绝、恶意、故意、刁难、报复、个人行为、威胁……

替代词集扩展。对种子词集经过 Word2Vec 工具作词向量处理后，结合 Synonyms 中文近义词查找工具，产生种子词集的替代词集，如“劝诫”的替代词有：告诫、责备等。实验中共取到 506 个词作为扩展的替代词，如表 4，其中 seed_T 替代词集（seed_T_synonym，下同）271 个，seed_F 替代词集（seed_F_synonym，下同）235 个：

表 4 替代词集词汇部分示例

类型	词集示例
seed_T_synonym	起诉、批捕、受理、裁定、调查结果、进行调查、调查报告、深入调查、书面形式、当面、劝诫、告诫……
seed_F_synonym	不能、没有、不会绝不、计入、扣除、算入、扣减、考核、绩效考核、奖惩、绩效评价、不计、数等、上列、所获、纳入……

4.3 犯罪线索筛查结果

犯罪线索筛查实验中使用未参与模型训练的 1 050 条数据，其中普通事件属性数据（F 类数据）1 000 条，疑似犯罪线索属性数据（T 类数据）50 条。根据本文所述基于语义的筛查方法对样本数据进行犯罪线索筛查积分运算，得到 F 类数据 997 条、T 类数据 53 条；经与实际数据比对，实际为 T 类且判定为 T 类的数据有 43 条，结果统计指标如表 5 所示。由于 T 类数据占全部待筛查数据比例较低，实验期待较高的结果召回率。从实验结果的指标来看，召回率 86%，精确率 81.13%，可以认为本文所述基于关键词挖掘的积分筛查模型在对热线文本数据中犯罪线索筛查时达到了预期的效果。

表 5 犯罪线索筛查方法结果指标统计

指标	准确率	精确率	召回率	漏报率
数值/%	98.38	81.13	86.00	14.00

5 结论

对热线数据中的犯罪信息做到有理、有据、科学的抽取是执法部门处理文本信息数据、确定犯罪线索的重要环节。本文提出了一种基于关键词挖掘的热线文本数据中犯罪线索自动化筛查方法，首先通过词向量模型及文本相似度计算等方法建立专业词库，然后设计了基于专业词库的犯罪线索积分筛查模型，并以济南市热线文本数据进行实证分析。经过与数据实际情况比对，该方法能够对济南市热线文本数据中的犯罪信息具体性识别并实现犯罪线索有效地筛查，使得犯罪线索筛查工作更加信息化和科学化。同时，该方法也适用于其他公安业务中文本数据目标信息识别及数据筛查，如舆情监测业务。本文也存在一定的局限，如在专业词库构建方面，词向量模型训练时需要一定数量的经验知识词汇及已知目标数据样本用于构建专业词库；在线索筛查算法方面，未来可引

入基于 doc2vec 的段落向量模型的文本分类方法，结合本文所述专业词库做定性加权分析。

参考文献：

[1] 王勇 . 大数据在我国食药智慧监管中的应用 [J]. 中国食品药品监管 , 2018(5): 44-47.
[2] 袁猛 , 刘文杰 , 胡建华 , 等 . "昆仑 2020" : 全方位构筑食药环安全防线 [J]. 人民公安 , 2020(16): 30-33.
[3] 徐建民 , 王金花 , 马伟瑜 . 利用本体关联度改进的 TF-IDF 特征词提取方法 [J]. 情报科学 , 2011, 29(2): 279-283.
[4] 彭云 , 万常选 , 江腾蛟 , 等 . 基于语义约束 LDA 的商品特征和情感词提取 [J]. 软件学报 , 2017, 28(3): 676-693.
[5] 刘耕 , 方勇 , 刘嘉勇 . 基于关联词和扩展规则的敏感词库设计 [J]. 四川大学学报 (自然科学版), 2009, 46(3): 667-671.
[6] 刘亚桥 , 陆向艳 , 邓凯凯 , 等 . 摄影领域评论情感词典构建方法 [J]. 计算机工程与设计 , 2019, 40(10): 3037-3042.
[7] 谭敏博 . 基于知识图谱的谷类作物病害识别及个性化推送研究 [D]. 长沙 : 湖南农业大学 , 2018.
[8] 夏松 , 林荣蓉 , 刘勘 . 网络谣言敏感词库的构建研究——以新浪微博谣言为例 [J]. 知识管理论坛 , 2019, 4(5): 267-275.
[9] 唐晓波 , 高和璇 . 基于关键词词向量特征扩展的健康问句分类研究 [J]. 数据分析与知识发现 , 2020, 4(7): 66-75.
[10] 姜天宇 , 王苏 , 徐伟 . 基于朴素贝叶斯的中文文本分类 [J]. 电脑知识与技术 , 2019, 15(23): 253-254, 263.
[11] 吴绍忠 . 重点人员积分预警模型建设基础问题研究 [J]. 中国人民公安大学学报 (自然科学版), 2012, 18(2): 76-79.
[12] 涂铭 , 刘祥 , 刘树春 . Python 自然语言处理实战核心技术与算法 [M]. 北京 : 机械工业出版社 , 2021:120, 129.
[13] 严红 . 词向量发展综述 [J]. 现代计算机 (专业版), 2019(8): 50-52.
[14] CHEN K J, MA W Y. Unknown word extraction for Chinese documents[C]// Proceedings of international conference on DBLP. Taipei: Morgan Kaufmann Publishers, 2002:169-175.
[15] PEDERSEN T, KULKARNI A. Identifying similar words and contexts in natural language with sense clusters[C]// Proceedings of the 20th national conference on artificial intelligence. Pittsburgh: AAAI Press, 2010:1694-1695.
[16] NEVIAROUSKAYA A, PRENDINGER H, ISHIZUKAM. SentiFul: a lexicon for sentiment analysis[J].IEEE transactions on affective computing, 2011, 2(1): 22-36.

作者贡献说明：
甄沐华：设计研究方法，完成实验，起草论文，修改论文与定稿；
陈　鹏：提出研究思路，修改论文；
王　坤：提供数据，提出研究问题；
范子杨：采集数据，进行实验；
王　者：采集数据，进行实验。

Research on Hotline Text Data Crime Clue Screening Method based on Keyword Mining

Zhen Muhua[1] Chen Peng[1] Wang Kun[2] Fan Ziyang[1] Wang Zhe[1]
[1]School for Informatics and Cyber Security, People's Public Security University of China, Beijing 100038
[2]Jinan Public Security Bureau, Jinan 250099

Abstract: [Purpose/Significance] Aiming at the problem of insufficient information analysis ability in the current public security business about identification and screening of crime clues in hotline texts, a method of hotline text data crime clue screening based on keyword mining is proposed to help business departments improve relevant intelligence and judgment **[Method/Process]** Considering that algorithms such as automatic text classification are subject to the problem of sample size, this paper firstly identified the key information of the known attribute data and established a seed lexicon, and then used Word2Vec to expand the seed vocabulary from the perspectives of similar words and synonym words to form a professional thesaurus, and finally used a semantics-based integral screening model to screen criminal clues in the hotline text data. **[Result/Conclusion]** This paper conducted a crime clue screening experiment on 1 050 priori hotline text data in Jinan City. After actual comparison and index analysis, the recall rate reached 86%. The specific identification of crime information in the text data of the city hotline achieved the expected effect and realized the effective screening of crime clues.

Keywords: hotline text　professional thesaurus　text similarity　crime clue screening